e-Research Infrastructures and Scientific Communication

Ralph Schroeder
Oxford Internet Institute University of Oxford

Jennifer A. deBeer
Oxford Internet Institute University of Oxford

Jenny Fry
Oxford Internet Institute University of Oxford

E-RESEARCH INFRASTRUCTURES AND SCIENTIFIC COMMUNICATION

Ralph Schroeder, Jenny Fry and Jennifer A. De Beer
Oxford Internet Institute
University of Oxford
1 St.Giles, Oxford, OX1 3JS, UK.
ralph.schroeder@oii.ox.ac.uk; jenny.fry@oii.ox.ac.uk; jennifer.debeer@oii.ox.ac.uk

Background

Infrastructures for e-Research are playing an increasing role within the rapidly growing domain of online scientific and scholarly communication. These infrastructures consist of networks of tools and data that are shared by communities of researchers. Current advanced technology developments in e-science raise the question as to whether infrastructures for e-science will become a small niche within the rapidly growing domain of online scientific communication - or if these systems, consisting of research instruments and scientific communication, will coalesce into a more broadly integrated system of knowledge production, dissemination and access. To be sure, the networks of data, tools and outputs that constitute current infrastructure developments, such as standards, ontologies, databases and e-print archives, and overlay journals, not to mention other such future scholarly services, are developing into a complex new system of knowledge production. Whether this new system can be said to constitute a movement towards a greater integration of scientific work, tools and resources or a more fragmented landscape of knowledge production and dissemination remains to be seen.

There are several challenges in these emerging infrastructures: One is the extent to which both institutional and epistemic (Knorr-Cetina, 1999) practices and policies promote 'openness', also often referred to as 'open access'. Such practices and policies are being developed within e-research, yet the provision of and experiences with developing shared access to e-research tools and infrastructures has been varied, and this is an especially important issue for developing countries. Another challenge lies in untangling the interrelated social, institutional and technical networks of this infrastructure in order to understand the relation between technical and policy issues for analytic purposes. The analytical challenge is partly due to the division of labour between the social science disciplines which deal with these new systems: the sociology of science and technology studies has focused on tools and how they are coupled to knowledge creation, whilst information science has focused on resources and how they relate to knowledge dissemination and access. Tools are closely coupled to data (primary resources) and methods and are the means by which data is processed and manipulated whereas secondary resources are the means by which data (or other epistemic objects (Rheinberger, 1997) are collected, organized and accessed. If we use structural bioinformatics to illustrate this distinction: primary resources are the large scale databanks, such as the *Protein Data Bank*, which are at the core of genomics research, tools are the algorithms that enable visualizations of the data and structure comparisons, and secondary resources are the semantic classificatory devices such as ontologies and metadata repositories (Zhang, Veretnik, and Bourne, 2005). As we shall see research enabled by advanced computing blurs traditional distinctions between primary resource, tools, and secondary resources in certain ways and so we need to draw on understanding and perspectives relating to both creation and dissemination of knowledge in order to follow emergent infrastructural challenges.

The paper is informed by a range of ongoing thematic studies related to e-Research, which include 'open science', 'sustainability' and 'e-infrastructure in developing countries'. These studies are based on interview, survey, and documentary evidence. We do not attempt to present an exhaustive inventory of current e-Research activities, open access policies, or initiatives to develop infrastructures to enable scientific communication and collaboration, as these are currently too fluid and multidimensional to allow for a comprehensive analysis. Rather, the paper aims to highlight some of the main challenges to emerge at the intersection where openness, e-Research, and e-infrastructures come together. To do this, we begin by outlining some recent changes that e-Research has introduced into the scientific and scholarly communication system. Next, we will give an account of e-Research systems in terms of their constituent

parts to get a sense of the layers and dynamics involved. Against this background, we can begin to see how access to and input to these systems can be problematic, particularly for developing societies. After sketching some of the ongoing initiatives to do this, we return to our main question: the outlook for the impact of e-Research on scientific and scholarly communication, particularly in the developing world.

Emergent Patterns in Scientific and Scholarly Communication

Currently, there is something akin to a paradigm shift taking place in scientific and scholarly communication. The intensity and impact of this shift differs across the heterogeneous fields of research that constitute the sciences, social sciences and arts and humanities. In areas such as biomedical research, genomics, computer science and particle physics, for example, the processing, storage and dissemination of 'data' is gaining importance, though practices with regard to 'openness' varies across disciplines. Networked databases, such as the RCSB (Research Collaboratory for Structural Bioinformatics) Protein Data Bank, have already become a core component of the information infrastructure in bioinformatics. Simulated, synthetic and in-silico data are playing an increasing central role in a number of other disciplines. For example, the British Atmospheric Data Center now have an archiving policy for simulated data (deterministic predictions (or hindcasts) based on algorithmic models as well as statistical analyses or composites of either or both of simulations and real data[1]). Born digital data and the algorithms associated with them are increasingly becoming recognized as valid scientific outputs by the institutions that govern academic research.

At the same time, however, traditional gatekeepers play a differential role across disciplinary communities. Some disciplines rely on the traditional system of publishing and peer review more than others – depending on factors such as the nature of scholarly recognition, intellectual pluralism, and certainty in research techniques and outcomes. In those fields where informal communication has historically played a critical role in establishing priority over ideas, the notion of 'open science' and the sharing of data may have more valence than in areas where formal communication plays a more central role in the dissemination of ideas. Disciplines that rely more on formal communication tend to be characterized by less-densely populated research niches, are more intellectually pluralistic and tend to use monograph style modes of communication to convey ideas (Becher and Trowler, 2001). Studies of patterns of computer-mediated communication within disciplines have shown that those disciplines that rely more on rapid informal communication, such as particle physics, are more likely to incorporate the internet into their knowledge dissemination practices. Consequently, we can observe the emergence of a fragmented communication system in relation to e-Research.

Some disciplinary communities, particularly those in the biomedical sciences, are advocating 'open science' and promoting it through open access archives and journals through initiatives such as the *NeuroCommons*[2]. One of the implications of these initiatives is that the dissemination of scientific research is more closely coupled to the provision of access to raw data, as is the case with overlay journals. In the UK, Europe and U.S. there have been recent trends in making the submission of datasets mandatory for publicly funded research and this has led to the establishment of funding agency data centres such as the UK Data Archive funded by the ESRC (Economic and Social Research Council) and the British Atmospheric Data Centre funded by NERC (Natural Environment Research Council). These data centres have various approaches to and mechanisms for governing 'openness'. At the 'closed' end of the spectrum are what Lane (2005) has coined 'data enclaves' in the social sciences whereby access to samples of data is controlled through End User Licences and other Special Licences, and access to larger samples of micro-social data is more stringently controlled. An example of this is the UK-based Samples of Anonymised Records (3% of the Census' UK) managed by the Cathie Marsh Centre for Census and Survey Research, whereby data disaggregated to the level of individuals and households are managed by the Office of National Statistics and can only be accessed via a limited number of physical locations in the UK.

In some disciplines 'open science' initiatives have led to collaboration with traditional information gatekeepers and the establishment of secondary resources such as *PubMed Central* in the biomedical

[1] Anne De Rudder, Jamie Kettleborough, Bryan Lawrence and Kevin Marsh "Archiving of Simulations within the NERC Data Management Framework: BADC Policy and Guidelines".

[2] See the "Background Briefing" on http://sciencecommons.org/projects/data/background.html

sciences, for example. In others, a hybrid system of 'overlay' journals connected to the cited archived datasets are emerging as is the case with the British Atmospheric Data Centre, and yet further approaches are developing alternative models of dissemination that by-pass traditional models of scientific and scholarly communication altogether. The Comb-e-Chem project, for example, is developing a 'publication@source' model that aims to establish a complete end-to-end connection between the results obtained at the laboratory bench and the final published analyses.

Regardless of the model being developed, these approaches are leading to a scholarly communication system based on the disaggregation of traditional scientific outputs, such as the scientific article, which can be linked to a 'unit-based' system of dissemination more closely aligned with the process of knowledge creation. Such 'units' could, for example, consist of data, images, simulations, software or preprints. Further, these outputs may then become identifiable in different ways, such as by means of digital object identifiers or classification schema for these objects – or it may simply be that outputs are now accessed or become accessible via internet search engines. Moreover, the possibility of measuring use of online units can serve as proxy for the calculation of the research impact of a work. In fact, whether or not a dataset has been cited in a published peer-reviewed journal can determine if it is deemed of significant value to archive.

The system of scientific or scholarly communication is thus undergoing changes which vary across disciplines. Nevertheless, this fragmented system affects formal and informal communication, data sharing and dissemination, gatekeeping and outputs. These changes also affect the 'openness' of access. There is thus an ongoing diversification of practices related to e-Research. Before we spell out the implications of these practices for 'open science' and for the developing world, however, it will be useful to get an overview of the system of e-Research – beyond scientific communication - and its component parts.

Levels and Components of Infrastructures and e-Research

To consider open access, it is important to have a comprehensive picture of all the levels and components of e-Research, but also to recognize that they may be subject to quite different dynamics. At the most general 'infrastructural' level, e-Research is subject to research policy, which is currently undergoing changes in the light of broader world-wide developments, for example in intellectual property regimes (Schroeder, 2007). At the most concrete and specific level, there are the everyday practices of researchers in relation to their e-Research work; accessing documents and data, sharing instruments and using repositories in distributed teams, searching for information and the like. In between the two are the various organizational and technical forces shaping e-Research, such as the agreements governing collaboration within and between institutions, the architecture of information stores, and the licenses and standards for middleware (David and Spence, 2004).

Openness is partly to do with legal restrictions or the absence thereof (Burk, 2007), but apart from this, openness also pertains to how these different levels and components interrelate. In the latter sense, openness signifies that the various parts of the electronic infrastructure and the tools connected to it should be able to interrelate in a flexible and seamless way. This, however, is difficult to achieve in practice, not simply because there are legacy systems and components, but also because there are continuous refinements and novel elements to the various parts which require alignment and updating throughout the system.

Thus, while we find that there is widespread support in principle for openness among researchers and policymakers, there is limited awareness of the complexities of how it needs to be implemented (David, den Besten and Schroeder, 2006). This should come as no surprise since the leaders of e-Research projects, for example, cannot be expected to be familiar with the intellectual property rights policies of their universities (arguably, they *should* be, but our interviews show that they have at best a cursory understanding). Conversely, policymakers specializing in intellectual property cannot be expected to know technical intricacies such as whether it is possible to combine different types of software that include components developed under different open source licenses.

What is to be done? It is hard not to embrace the principles that have emerged within the sociology of science and technology that have been said to respond to the challenges of these new complex systems: bottom-up generativity instead of top-down control, heterogeneity instead of standardization and homogeneity, and of course openness and flexibility instead of closed and centralized systems (see Hughes, 1998). At the same time, it is difficult to see how coordinated drives towards particular governance regimes, lock-ins to particular systems or certain systems becoming the single dominant standard, and grafting of parts into a congealed whole with variable degrees of fixity – can be avoided.

The paradox is that although scientific and scholarly communication is open in principle, in the sense that it is subject to continual refinement (Fuchs, 2002; Becher and Trowler, 2001), e-Research consists of networks and systems that are still in an early stage of development and the forms of openness are still fluid. Yet unlike science-in-the-making which continually incorporates consensus and moves on to new territory, the technology-in-the-making of e-Research aims to create shared systems and resources that are able to support research collaboration over the longer term – in short, it aims to create stable technological and social structures rather than being open to constant flux. This contrast may be overdrawn since there are some forms of knowledge that are rather stable - and there are technological systems that change rapidly. Nevertheless, openness in e-Research remains entangled in a web of more and less congealed strictures. This point could also be made in a different way by reference to the distinction between tools and resources that was mentioned earlier: instead of distinguishing between fluid knowledge as against static structures, it may be useful to distinguish between tools and resources.

Tools versus Resources

Tools are the means for manipulating information and data, and these nowadays consist of software as well as of computer processing and storage capabilities. Resources, on the other hand, consist of the information that is accessed for research, and consist, in addition to traditional publications, of digital archives and databases. Both tools and resources can be found on the various levels of these e-Research systems.

Not only are they spread across the e-Research system, but it may also be that the networks of data and outputs that constitute current infrastructure developments – which include the development of standards, ontologies, searchable databases and e-print archives - blur the traditional distinction between tools and resources in science. If researchers can, for example, discover new knowledge by linking data, by improving the means by which large amounts of information can be processed, or by developing schemata which allow information to be searched in new ways – then perhaps this blurs the distinction between the creation (by means of a tool) and dissemination (as a resource) of knowledge.

Tools and resources are both subject to incremental improvement in e-Research; neither tools nor resources are completely fluid or static. And both can be open or "closed"; in terms of input and output (or contributions and access) in the case of resources, and in terms of development (including standards) and access in the case of tools. However (and this will become important later), tools in e-Research are often modeled on open source development and primarily require skills, whereas resources, even if there is an impetus towards open access, require costly networks and access to expensive-to-maintain publications and databases.

In any event, both tools and resources (in both senses mentioned earlier, the primary resources of data and secondary resources of how they are accessed, collected and organized) in the case of e-Research are also parts of larger institutional and organizational infrastructures which require funding and skills and they grow, diffuse and develop a sustainable momentum as parts of these larger systems. This is why, even if we can assume that there is considerable momentum of these systems in developed societies, we cannot assume that they can be integrated in the Global South.

Shared access initiatives in - and for - developing countries

If we consider e-Research infrastructures as comprising both the networks and technologies underpinning science systems, as well as the scientific communication which happens on top of or through them, we can broadly identify two categories of initiatives. The first category is those which address the participation of

developing countries in what are variously referred to as cyberinfrastructure, e-Science, or e-infrastructure initiatives. These differences in nomenclature attest more to different etymologies than that they refer to distinct concepts: Cyberinfrastructure is the term used in the United States, e-Science is the United Kingdom´s neologism, and e-infrastructure that of the European Union. These terms are all understood to refer to the networks, technologies, and organisational setups which, in particular, support large-scale and geographically diffuse research collaboration. Tacitly, the use of these terms serves to connote an opening up of, and openness within, scientific research practice (Schroeder & Fry, 2007) regardless of discipline.

The second category of initiatives consists of the dissemination of previously mentioned disaggregated traditional scientific output, whether these take the form of data and/or traditional research publications in their various stages of maturity, complemented by image, audio or video captures. At this point we can map the initiatives for both of the network-level and content dissemination categories within the developing world or Global South, broadly comprising Africa, Latin America, India, and China. Whereas the notion of e-Infrastructure at European level conjures up an image of high-speed networks, when we cast our gaze in the direction of the developing world we are often still confronted with the lo-fi version of the conditions in the developed world. Participation by developing countries in high-performance or GRID initiatives is still the exception rather than the rule. Many in the developing world have only recently become exposed to improved bandwidth[3] via interconnection with European or United States networks, these being either GEANT and ABILENE, respectively. Much like the original GÉANT, its successor, GÉANT2 runs under the auspices of the European Union´s DANTE (Delivery of Advanced Technology to Europe) project, and in turn, has links projecting into Latin America (RedCLARA); the Asia-Pacific region (TEIN2); the southern and eastern Mediterranean countries (EUMEDCONNECT) and sometimes directly to developing countries, sometimes bypassing regional/consortial interconnections, as is the case for South Africa, India, and China. What we see is that slowly, sometimes surely, various parts of the Global South start to interconnect with high-speed networks.

More particularly, within and across Africa, various research and education networks (RENs) have come to fruition largely during the past two to three years, with these taking the form of regional REN consortia, consisting in turn of national RENs. The two notable African initiatives are the Ubuntunet Alliance[4], and the African Virtual University Bandwidth Consortium[5]. Both initiatives are multi-country university-led efforts at collective bargaining for cheaper bandwidth, whilst also focusing on the set-up and expansion of physical network infrastructure. These initiatives are relatively new, reported on at a meeting in December 2006[6], and as such have only recently joined the global forum. At the same time, though an ideal, it cannot be guaranteed that all of the African RENs will interconnect with such global initiatives, given that sustained
e-Infrastructure investment would need to be made, and often other basic infrastructures, such as water or energy supply, are in need of establishment[7].

It seems then that there is at best a mixed picture in terms of how developing countries, despite these interconnections, still lag behind when it comes to being at the cutting-edge in the use of truly advanced networks. We see further that the use of the term Global South is misleading since regions vary in their levels of participation. We can say, for instance, that Latin America runs ahead of Africa, and so to use such blanket terminology does not suffice. Projects such as BELIEF[8] (Bringing Europe´s eLectronic Infrastructures to Expanding Frontiers) and 6DISS, are European Union projects aimed at taking networks to the next level in developing countries. They are thus efforts at bringing the next layer of network infrastructure to the developing world, i.a. distributing IPv6 (via 6DISS), or planning for next-level uses of

[3] Which is not to say that connectivity was wholly lacking before in all of these regions; rather that the data throughput, where extant, had been lower previously.
[4] http://www.ubuntunet.net/
[5] http:// www.avu.org/
[6] http://www.wideopenaccess.net/2006/
[7] See the European Commission´s EU-Africa Partnership on Infrastructure (COM(2006) 376 final, 13 July 2006)

networks as they evolve, e.g. supercomputing. From the aforementioned, it can only be concluded that the rate of participation by Global South countries is at best staggered, when at all existent.

If we consider infrastructures in terms of the category of disseminating scientific research output, there is the now-familiar two-pronged schema of Open Access journals and institutional repositories for the distribution of digital versions of the traditional research paper (Jeffery, 2006). Yet, as indicated above, we should not restrict our sights merely to these traditional uses. In the developing world the primary initiatives to date for access to journal content by developing country scientists and researchers have been to provide research institutions in a predefined list of developing countries (according to GDP per capita) free access to the full-text of journals. Note that on the basis of the GDP-per capita entry-gate, some developing countries are therefore excluded from the benefits of such a scheme. Some other initiatives aim to improve journal publication practices in developing countries, whilst others make the abstracts available online for free, of a stable of journals, with the option of document request and delivery to end-users.

While access to journal content can be seen as efforts from the developed world to assist the developing world in having greater access to research publications, the creation of institutional repositories can be read as the extent to which developing country research institutions and universities are prepared or capable of helping themselves with research dissemination, since the different types of software[9] for creating institutional repositories are freely available. What may be lacking however, though not exclusively so, is either a lack of capacity to set up the requisite systems or lack of institutional will to do so. The OpenDOAR[10] human-compiled directory of Open Access institutional repositories indicates that, of the 627 organisations[11] indexed in the directory, a mere 1% (6 sites) can be attributed to the African continent, 1% (4 sites) in Central America, 4% (22) in South America, and 6% (40) in Asia.

Since having access to infrastructure is a requirement for access to research content; and considering further that other types of latterday collaborative network-enabled scientific research cannot be done without such connectivity, it is worrying to note that as much as e-Research takes off in the developed world, a parallel, slow and uneven adoption, or even at times non-adoption of such practices for the developing world can be predicted. The challenge here then is consideration of how the developing world may be kept in line with e-Research developments in the developed world.

Conclusion

The growing diversity of practices in scientific and scholarly communication can be related to the challenges of emerging e-infrastructures, which include the tension between local work practices and goals at the level of individual projects and the necessity to coordinate contributions at a macro-level, perhaps even beyond specific knowledge domains, towards a sustainable infrastructure.

As we have seen, e-Research systems add a layer of complexity to an already complex communication system. Making these systems extend to the developing world, including openness not only in terms of open access, but also openness practices ranging from sharing infrastructure capacities and linking on various technical and organizational levels, to uses of common tools and resources, therefore involves a range of issues. Openness cuts across these inasmuch as it enables flexible and interoperable systems and components, but it also requires tools and resources and the sustainability of both under conditions when the lines between them are blurring. Thus we need to combine a variety of perspectives, including the sociology of science and technology, information science, and research policy, to recognize the limits of a congealing system, and the conditions for making it less so and more open and encompassing. Whilst also bearing in mind that disciplines such as Development Studies, or Area Studies covering the developing world, may make valuable contributions in furthering thinking on how the developing world may become and remain active in such rapidly-advancing scientific arenas.

---

[9] In the main, there are two packages, Eprints, Dspace.

[10] http://www.opendoar.org/

[11] At the time of writing there are more than 800 institutional repository (IR) installations worldwide, but some organisations may house more than one archive, hence the more reliable means of tallying continent-wide installations is by tallying organisations.

References

Becher, T. and Trowler, P. (2001) 2nd ed. *Academic Tribes and Territories: Intellectual Inquiry and the Culture of Disciplines*. Milton Keynes: Open University Press.

Burk, D. (2007) Intellectual Property in the Context of e-Science, *Journal of Computer-Mediated Communication,* Volume 12, Issue 2, January.

David, P. A., & Spence M. (2004) Towards a cyberinfrastructure for enhanced scientific collaboration. *OII Research Report No. 4*. Oxford: Oxford Internet Institute, University of Oxford. Available at: http://www.oii.ox.ac.uk/microsites/oess/papers.cfm. Accessed: 28th September 2006.

David, P., den Besten, M., and Schroeder, R. (2006) How Open is e-Science? *Proceedings of IEEE e-Science*, Amsterdam, December 4-6.

Fuchs, S. (2002) What Makes Sciences Scientific?, in Jonathan Turner (ed.), *Handbook of Sociological Theory*, New York: Kluwer Academic/Plenum Publishers, pp.21-35.

Hughes, T. (1998) *Rescuing Prometheus.* New York: Pantheon Books.

Jeffery, K. (2006) Open access: An introduction. *ERCIM* News (January). European Research Consortium for Informatics and Mathematics. Available at: http://www.ercim.org/publication/Ercim_News/enw64/jeffery.html. Accessed: 10th April 2007.

Knorr-Cetina, K. (1999) *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, Massachusetts, Harvard University Press.

Lane, J. (2005) *Optimizing the Use of Micro-Data: An Overview of the Issues*. Available at: http://ssrn.com/abstract=807624. Accessed: 24th April 2007.

Rheinberger, H-J. (1997). *Toward History of Epistemic Things: Synthesizing Proteins in the Test Tube.* Stanford, CA, Stanford University Press.

Schroeder, Ralph. 2007. e-Research Infrastructures and Open Science: Towards a New System of Knowledge Production? *Prometheus*, vol.25, no.1, pp.1-17.

Schroeder, Ralph and Fry, Jenny. 2007. Social Science Approaches to e-Science. *Journal of Computer-Mediated Communication.* Volume 12, Issue 2, January 2007.

Zhang, Q., Veretnik, S., and Bourne, P. (2005) Overview of Structural Bioinformatics. *In:* Yi-Ping Phoebe Chen (Ed). *Bioinformatics Technologies*. Springer-Verlag, Germany.